Ubiquitous and Mobile Computing
CS 528: Unsupervised Speaker Counter with Smartphones

Xuanyu Li

Computer Science Dept.

Worcester Polytechnic Institute (WPI)

Introduction

- Conversation is very important !
  - Most direct form of social interactions
- Relevant researches
  - Speaker Identification
  - Characterization of social settings
- BUT what might be overlooked ???

Introduction

- Speak counter: measurement of number of people in a conversation
- App name: crowd++
- Motivation?

  Social hotspot

  Social diary

  LAST BUT NOT LEAST ?

    Participation Estimation (class participation)

Challenges

- Location (pocket or bag)
- hardware constraints
- noise polluting

System Design

First step: Speech detection

- Target: filter out silence periods and background noise
- Divide speech into segments (3s/segment)
- 3s? Provides good trade-off between inference delay and accuracy
- Tradition: energy-based voice data detection (unsuitable for mobile device)
- Crowd++: Pitch

System Design

- Second step: Feature Extraction
  - Precondition: filtered out non-speech/background noise
  - Postcondition: extracted features can effectively distinguish speakers
  - The Less overlap, the better

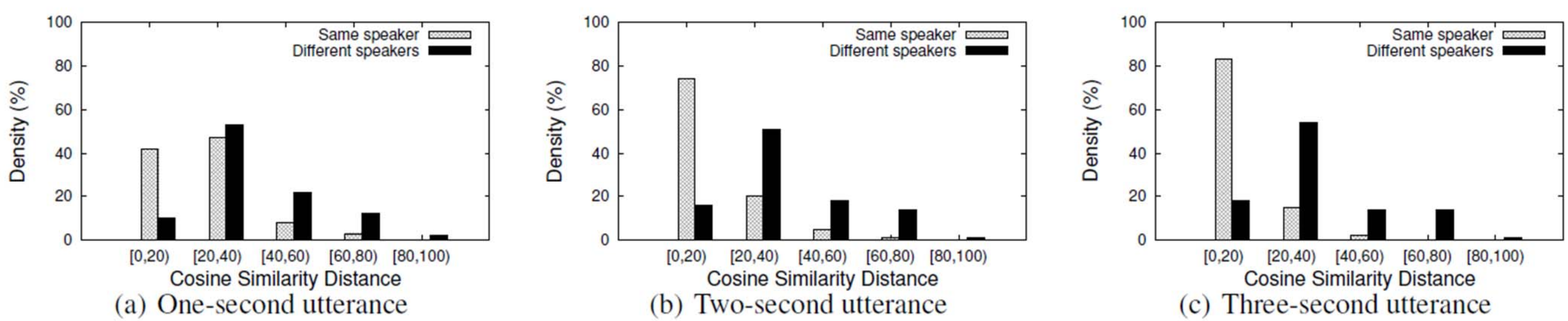

Figure 2. Cosine similarity distance demonstrates better speaker distinguishing capabilities with longer utterance.

System Design

- Counting Engines
  - Counting algorithm
    - Traditional: hierarchical clustering
      - Compares each segment with the other, thus runs in O(n^2) time ({S1, S2, S3, …… , Sn})
    - Crowd++: forward clustering
      - Compares adjacent segments and merge the similar ones, runs in O(n) time ({((S1, S2), S3), S4 ……, Sn})

System Design

- If (S1 close to S2) {
  - merge(S1, S2) to S1;
  - compare S1 with S3;

  } else

  compare S2 with S3;

…… do above recursively until traverse is done

Evaluation

- Performance metrics:
  - Name : Error Count Distance
  - Definition: |C^ – C|
    - C^: estimated number by the app
    - C: real number of participants

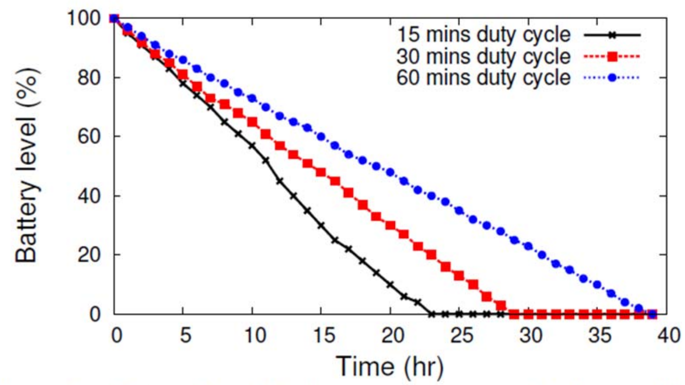

Figure 3. A duty-cycle of 15 mins guarantees a one day battery life for the Samsung Galaxy S2.

- Energy consumptions
  - Cycling: 5min recording + algorithm + sleep(T interval)
  - Lower bound performance (battery)
  - Mainly used in public location

Performance with a single group

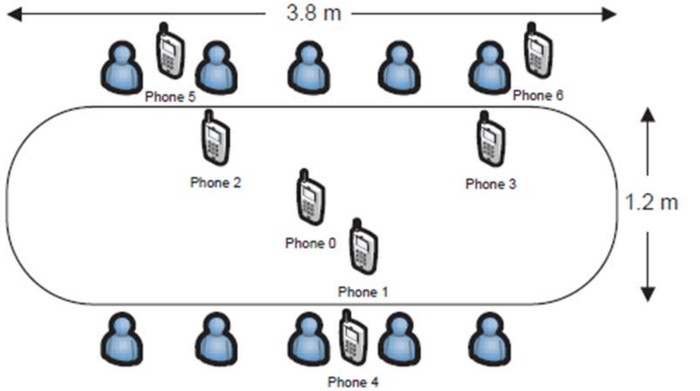

Figure 4. The phone placement in the benchmark experiments.

1. Phone 0-3 on the table

2. Phone 4-6 in users pocket

Conclusion:

- If on table, position does not matters much
- In pocket is not as accurate as on table

Performance with multiple groups

- For instance: Restaurant

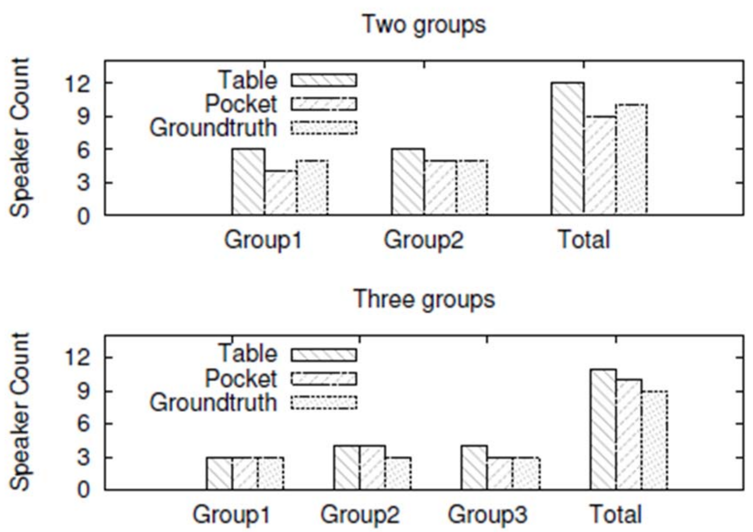

Figure 8. The phones inside the pockets present better counting results when multiple groups of speakers are co-located.

Something quite interesting is that ……

Possible explanation:

Pocket phone has better ability to filter out distant sound

Performance with various conversation parameters

- Audio Clip Duration (longer, better)
- Overlapping Percentage (No noticeable influence found)
- Utterance Length (0-3s fluctuate, >3s stable with error distance decreased to 1)

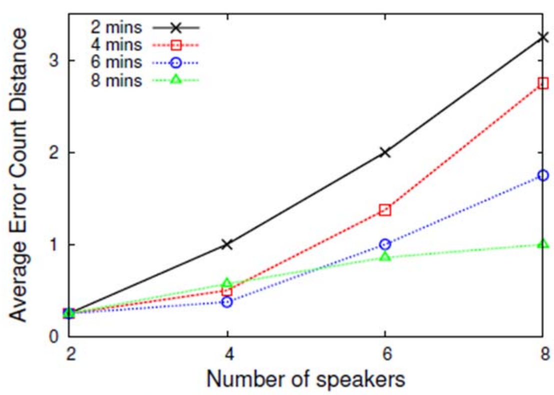

Figure 9. Eight-minute audioclips are sufficient to achieve an error count distance of 1.

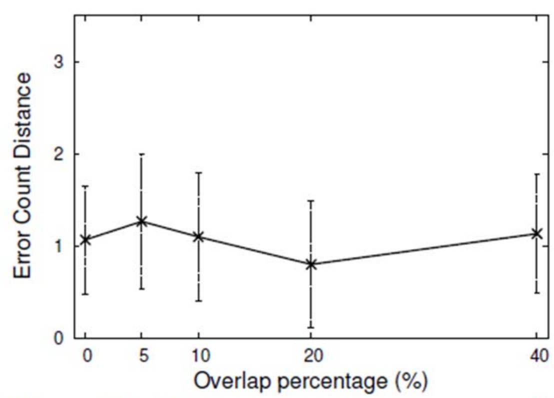

Figure 10. The average counting error distance is around 1 with up to 40% overlap.

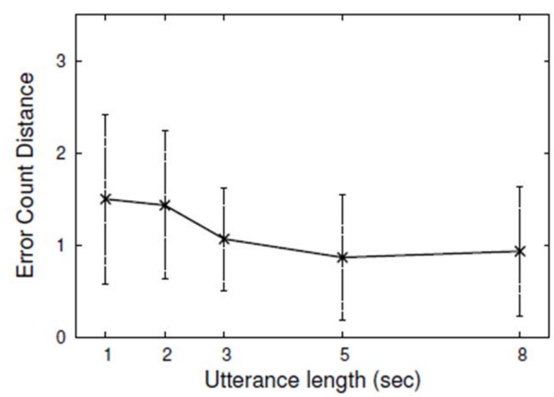

Figure 11. Longer utterance lengths lead to slightly better counting performance.

Privacy Concerns

- Speaker's identification is never revealed
  (extra algorithms)
- Data analysis is always performed locally in case of data leakage
- User has the option when to activate the application

Conclusion

- Unsupervised (no prior models, external hardware)
- No machine learning algorithms
- Totally local on device
- Great accuracy with low error distance
- Multiplatform support

References

1. Agneessens, A., Bisio, I., Lavagetto, F., Marchese, M., and Sciarrone, A. Speaker count application for smartphone platforms. In *Proc. of IEEE ISWPC* (2010).

2. Anguera Miro, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. Speaker diarization: A review of recent research. *IEEE Transaction on Audio, Speech and Language Processing 20*, 2 (2012).

3. Azizyan, M., Constandache, I., and Roy Choudhury, R. Surroundsense: mobile phone localization via ambience fingerprinting. In *Proc. of ACM MobiCom* (2009).

4. Baken, R. *Clinical measurement of speech and voice*. College-Hill Press, 1986.

5. Carey, M., and et al. Robust prosodic features for speaker identification. In *Proc. of ICSLP* (1996).

6. Cetin, O., and Schriberg, E. Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap. In *Proc. of IEEE ICASSP* (2006).

7. Chan, A. B., Liang, Z.-S., and Vasconcelos, N. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proc. of IEEE CVPR* (2008).

8. Cheveigné, A. D., and Kawahara, H. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America 111*, 4 (2002).

9. Choudhury, T., and Pentland, A. Sensing and modeling human networks using the sociometer. In *Proc. of IEEE ISWC* (2003).

- Thank you !